Esercizi - Analisi Predittiva - II

aa 2022/2023

Appunti per la Soluzione ad esercizi scelti

Esercizio 1

Si consideri il dataset `wbca` del pacchetto `faraway`. Il dataset contiene dati riguardo uno studio oncologico in cui si vuole poter individuare se un tumore è maligno o meno usando alcune caratteristiche delle cellule estratte usando un ago aspirato. Si veda anche `?faraway::wbca`.

1. Si esegua una prima stima in cui la variabile `Class`, cioè la variabile che indica la classificazione del tumore, dipende dalla variabile `Thick`: si produca un grafico che mostra come la probabilità che un tumore sia benigno dipende da `Thick`. Si crei inoltre un intervallo di confidenza al 96% per il parametro relativo alla variabile `Thick`.

2. Si stimi un modello in cui la la variabile `Class` dipende da tutte le altre variabili disponibili nel dataset. Si ottenga la devianza residua del modello e si verifichi se il modello ha un qualche valore predittivo. Si verifichi inoltre se il modello ha una miglior capacità predittiva del modello in cui solo la variabile `Thick` è usara come predittore.

3. Si usi la funzione `step` per costruire un modello in cui una sottoinsieme delle variabili viene usato. Si confrontino i modelli ottenuti quando vengono usati AIC o BIC come criteri per scegliere il sottoinsieme di variabili da usare come predittori nel modello. Se vi è qualche differenza tra i due modelli identificati usando i due criteri, si indichi il motivo alla base della differenza.

4. Usando il modello selezionato usando AIC si stimi la probabilità che due pazienti con le caratteristiche indicate nel dataset `nd` abbiano un tumore benigno (cioè un tumore con `Class = 1`):

   ```
   nd <- data.frame( Patient =  c("A","B"),
       Adhes = c(1,3), BNucl = c(1,3.5), Chrom = c(3,3.5), Epith = c(2,3.5),
       Mitos = c(1,1.6), NNucl = c(1,2.8), Thick = c(4,4.43), UShap = c(1,3.2),
       USize = c(1,3.14))
   ```

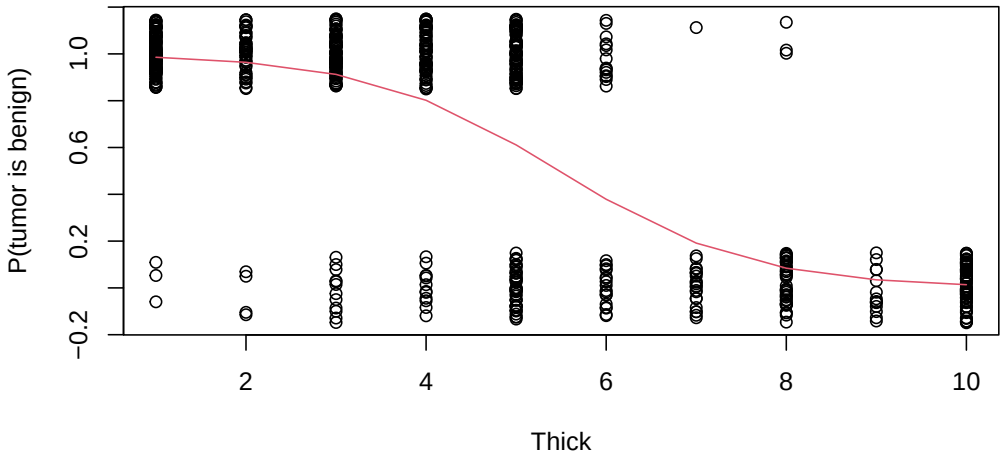
   Si fornisca una stima puntuale e una stima intervallare di questa probabilità usando un livello di confidenza del 96%. Si commenti l'ampiezza degli intervalli identificati.

Soluzione Esercizio 1

```
data(wbca,package="faraway")
```

1. ```
   tumor_thick <- glm(Class~Thick, data = wbca, family=binomial)
   plot(jitter(Class, amount = 0.15)~Thick, data = wbca, ylab = "P(tumor is benign)")
   lines(sort(wbca$Thick), fitted(tumor_thick)[order(wbca$Thick)],col=2)
   confint.default(tumor_thick, parm = "Thick", level = .96)
   ```
   ```

```
              2 %       98 %
Thick -1.102012 -0.7901991
```



2. *tumor_all <- glm(Class~., data = wbca, family=binomial)*
   deviance(tumor_all)

```
[1] 89.4642
```

anova(tumor_thick, tumor_all, test = "LRT")

```
Analysis of Deviance Table

Model 1: Class ~ Thick
Model 2: Class ~ Adhes + BNucl + Chrom + Epith + Mitos + NNucl + Thick +
    UShap + USize
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       679     451.69
2       671      89.46  8   362.23 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

AIC(tumor_thick, tumor_all)

```
            df      AIC
tumor_thick  2 455.6905
tumor_all   10 109.4642
```

La devianza residua per il modello `tumor_all` é 89.46: possiamo confrontare questo valore al valore della devianza residua per il modello `tumor_thick` (che è annidato in `tumor_all`) tramite una tavola

di analisi della devianza o possiamo fare un confronto basato su criteri di informazione quali AIC e BIC che danno una misura della bontà di adattemento generale. Vediamo che il modello più complesso risulta significativamente diverso da modello più semplice: alcuni dei predittori sono utili a spiegare la varibailità dei dati in aggiunta alla variabile `Thick`.

3. Se si usa BIC si trova un modello più parsimonioso: questo perché BIC una una penalizzazione più forte di AIC e quindi penalizza di più modelli con molti predittori.

```
tumor_null <- glm(Class~1, data = wbca, family=binomial)
tumor_selAIC <- step(tumor_all, direction = "both",
                     scope = list(lower=tumor_null, upper=tumor_all))
```

```
Start:  AIC=109.46
Class ~ Adhes + BNucl + Chrom + Epith + Mitos + NNucl + Thick +
    UShap + USize

        Df Deviance    AIC
- USize  1   89.523 107.52
- Epith  1   89.613 107.61
- UShap  1   90.627 108.63
<none>       89.464 109.46
- Mitos  1   93.551 111.55
- NNucl  1   95.204 113.20
- Adhes  1   98.844 116.84
- Chrom  1   99.841 117.84
- BNucl  1  109.000 127.00
- Thick  1  110.239 128.24

Step:  AIC=107.52
Class ~ Adhes + BNucl + Chrom + Epith + Mitos + NNucl + Thick +
    UShap

        Df Deviance    AIC
- Epith  1   89.662 105.66
- UShap  1   91.355 107.36
<none>       89.523 107.52
+ USize  1   89.464 109.46
- Mitos  1   93.552 109.55
- NNucl  1   95.231 111.23
- Adhes  1   99.042 115.04
- Chrom  1  100.153 116.15
- BNucl  1  109.064 125.06
- Thick  1  110.465 126.47

Step:  AIC=105.66
Class ~ Adhes + BNucl + Chrom + Mitos + NNucl + Thick + UShap
```

```
        Df Deviance    AIC
<none>       89.662 105.66
- UShap  1   91.884 105.88
+ Epith  1   89.523 107.52
+ USize  1   89.613 107.61
- Mitos  1   93.714 107.71
- NNucl  1   95.853 109.85
- Adhes  1  100.126 114.13
- Chrom  1  100.844 114.84
- BNucl  1  109.762 123.76
- Thick  1  110.632 124.63
```

```
 tumor_selBIC <- step(tumor_all, direction = "both", k = log(nrow(wbca)),
                      scope = list(lower=tumor_null, upper=tumor_all))
```

```
Start:  AIC=154.7
Class ~ Adhes + BNucl + Chrom + Epith + Mitos + NNucl + Thick +
    UShap + USize

        Df Deviance    AIC
- USize  1   89.523 148.24
- Epith  1   89.613 148.32
- UShap  1   90.627 149.34
- Mitos  1   93.551 152.26
- NNucl  1   95.204 153.92
<none>       89.464 154.70
- Adhes  1   98.844 157.56
- Chrom  1   99.841 158.55
- BNucl  1  109.000 167.71
- Thick  1  110.239 168.95

Step:  AIC=148.24
Class ~ Adhes + BNucl + Chrom + Epith + Mitos + NNucl + Thick +
    UShap

        Df Deviance    AIC
- Epith  1   89.662 141.85
- UShap  1   91.355 143.54
- Mitos  1   93.552 145.74
- NNucl  1   95.231 147.42
<none>       89.523 148.24
- Adhes  1   99.042 151.23
- Chrom  1  100.153 152.34
+ USize  1   89.464 154.70
- BNucl  1  109.064 161.25
- Thick  1  110.465 162.65
```

```
Step:  AIC=141.85
Class ~ Adhes + BNucl + Chrom + Mitos + NNucl + Thick + UShap

        Df Deviance    AIC
- UShap  1   91.884 137.55
- Mitos  1   93.714 139.38
- NNucl  1   95.853 141.52
<none>       89.662 141.85
- Adhes  1  100.126 145.79
- Chrom  1  100.844 146.51
+ Epith  1   89.523 148.24
+ USize  1   89.613 148.32
- BNucl  1  109.762 155.43
- Thick  1  110.632 156.30

Step:  AIC=137.55
Class ~ Adhes + BNucl + Chrom + Mitos + NNucl + Thick

        Df Deviance    AIC
- Mitos  1   96.494 135.63
<none>       91.884 137.55
+ UShap  1   89.662 141.85
- NNucl  1  103.711 142.85
+ USize  1   90.923 143.11
+ Epith  1   91.355 143.54
- Adhes  1  105.473 144.61
- Chrom  1  109.699 148.84
- BNucl  1  124.813 163.96
- Thick  1  130.842 169.98

Step:  AIC=135.64
Class ~ Adhes + BNucl + Chrom + NNucl + Thick

        Df Deviance    AIC
<none>       96.494 135.63
+ Mitos  1   91.884 137.55
+ UShap  1   93.714 139.38
+ USize  1   95.042 140.71
+ Epith  1   95.868 141.53
- Adhes  1  110.725 143.34
- NNucl  1  111.384 144.00
- Chrom  1  114.153 146.77
- BNucl  1  128.941 161.56
- Thick  1  149.705 182.32
```

4.
```
nd <- data.frame( Patient =  c("A","B"),
    Adhes = c(1,3), BNucl = c(1,3.5), Chrom = c(3,3.5), Epith = c(2,3.5),
    Mitos = c(1,1.6), NNucl = c(1,2.8), Thick = c(4,4.43), UShap = c(1,3.2),
    USize = c(1,3.14))
predict(tumor_selAIC, newdata = nd, type = "response")
```

```
        1         2
0.9921115 0.7245763
```

```
preds <- predict(tumor_selAIC, newdata = nd, type = "link", se.fit = TRUE)
pint <- cbind(tumor_selAIC$family$linkinv(preds$fit + qnorm(.02) * preds$se.fit),
              tumor_selAIC$family$linkinv(preds$fit + qnorm(.98) * preds$se.fit))
pint
```

```
       [,1]      [,2]
1 0.9744226 0.9975972
2 0.5663004 0.8412797
```

```
# ampiezza degli intervalli
pint[,2]-pint[,1]
```

```
         1          2
0.02317464 0.27497929
```

```
summary(wbca)
```

```
    Class            Adhes           BNucl           Chrom
 Min.   :0.0000   Min.   : 1.000   Min.   : 1.000   Min.   : 1.000
 1st Qu.:0.0000   1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.: 2.000
 Median :1.0000   Median : 1.000   Median : 1.000   Median : 3.000
 Mean   :0.6505   Mean   : 2.816   Mean   : 3.542   Mean   : 3.433
 3rd Qu.:1.0000   3rd Qu.: 4.000   3rd Qu.: 6.000   3rd Qu.: 5.000
 Max.   :1.0000   Max.   :10.000   Max.   :10.000   Max.   :10.000
    Epith            Mitos           NNucl           Thick
 Min.   : 1.000   Min.   : 1.000   Min.   : 1.000   Min.   : 1.000
 1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.: 2.000
 Median : 2.000   Median : 1.000   Median : 1.000   Median : 4.000
 Mean   : 3.231   Mean   : 1.604   Mean   : 2.859   Mean   : 4.436
 3rd Qu.: 4.000   3rd Qu.: 1.000   3rd Qu.: 4.000   3rd Qu.: 6.000
 Max.   :10.000   Max.   :10.000   Max.   :10.000   Max.   :10.000
    UShap            USize
 Min.   : 1.000   Min.   : 1.00
 1st Qu.: 1.000   1st Qu.: 1.00
 Median : 1.000   Median : 1.00
 Mean   : 3.204   Mean   : 3.14
 3rd Qu.: 5.000   3rd Qu.: 5.00
 Max.   :10.000   Max.   :10.00
```

Il primo paziente è un paziente molto più tipico del secondo, dato che i valori osservati per le diverse caratteristiche sono simili ai valori medi dei pazienti osservati nel campione. Di conserguenza la stima è meno incerta per il primo paziente.

Esercizio 2

Gli abitanti di sesso maschile dell'isola greca di Kalythos soffrono di una malattia congenita agli occhi, i cui effetti diventano più marcati in età avanzate. Su un campione di isolani di sesso maschile e di età diverse è stato contato il numero di individui ciechi. Il codice crea un dataset per i dati osservati creando una variabile per il numero di totale di uomini campionati per ogni età e una variabile per il numero di uomini ciechi individuati nel campione:

```
Kalythos <- data.frame(age = c(20,35,45,55,70),
                       total_sample = c(50,50,50,50,50),
                       n_blind = c(6,17,26,37,44))
```

1. Si stimi un modello che indaghi se la proporzione di persone con cecità nell'isola cambia in funzione dell'età degli individui. Si usi la funzione legame (link function) canonica.

2. Si crei un grafico che mostri la relazione stimata dal modello al punto precedente: si commenti come il modello stimato si adatta ai dati raccolti

3. Si calcoli un intervallo di confidenza della probabilità che ha un individuo nell'isola di essere cieco a 20, 50 e 70 anni. Si usi un livello di confidenza pari al 90%

4. Si proceda a fare un test per testare se il valore del coefficiente relativo al predittore `age` è uguale a 0.1. Si usi un livello di significatività del 10%.

5. Si delinei brevemente la base teorica usata per derivare il test svolto nel punto precedente, commentando la validità di tale base per l'applicazione al punto 4.

6. Si usino le due variabili specificate qui sotto come predittori in un modello che indaghi come la proporzione di persone con cecità nell'isola cambia in funzione dell'età degli individui. Si confrontino i valori stimati dei coefficienti: che interpretazione si può dare alla stima dell'intercetta nei diversi modelli?

   ```
   Kalythos$age_m20 <- Kalythos$age-20
   Kalythos$age_m45 <- Kalythos$age-45
   ```

Soluzione Esercizio 2

```
Kalythos <- data.frame(age = c(20,35,45,55,70),
                       total_sample = c(50,50,50,50,50),
                       n_blind = c(6,17,26,37,44))
```

1. ```
   Kalythos$n_healthy <- Kalythos$total_sample- Kalythos$n_blind
   f0 <- glm(cbind(n_blind, n_healthy) ~ age, data = Kalythos, family = binomial)
   summary(f0)
   ```
   ```

```
Call:
glm(formula = cbind(n_blind, n_healthy) ~ age, family = binomial,
    data = Kalythos)

Deviance Residuals:
      1        2        3        4        5
-0.1797   0.1157  -0.1182   0.3791  -0.3372

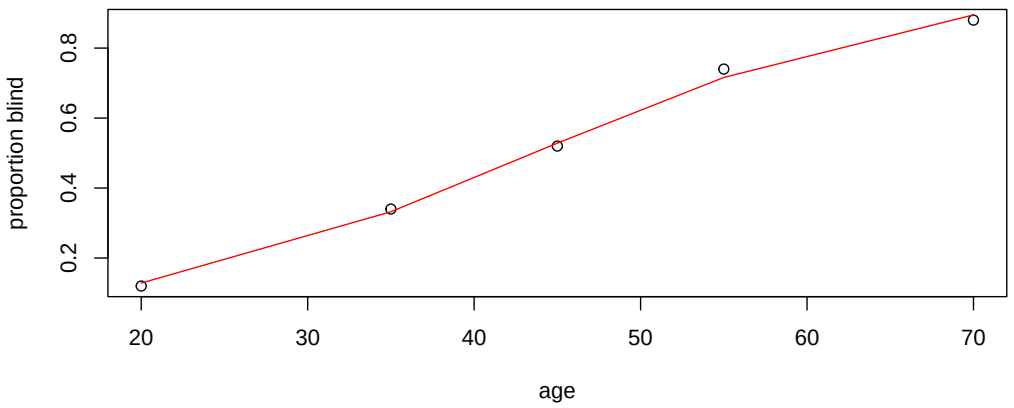
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.53778    0.50232  -7.043 1.88e-12 ***
age          0.08114    0.01082   7.498 6.47e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 82.14455  on 4  degrees of freedom
Residual deviance:  0.31707  on 3  degrees of freedom
AIC: 24.132

Number of Fisher Scoring iterations: 4
```

```
plot(Kalythos$age, Kalythos$n_blind/Kalythos$total_sample,
       ylab = "proportion blind", xlab = "age")
lines(Kalythos$age, f0$fitted.values, col="red")
```



2. Visivamente si nota che il modello si adatta piuttosto bene ai dati osservati: al crescere dell'età cresce la probabilità che le persone siano cieche.

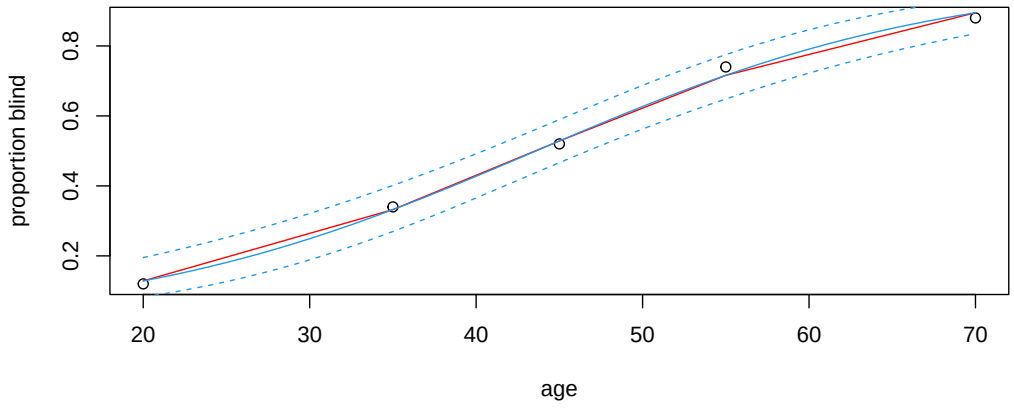
3. Deriviamo gli intervalli di confidenza

```
nd <- data.frame(age=seq(20,70,by=1))
pred <- predict(f0, newdata = nd, se.fit=TRUE, type="link")
plot(Kalythos$age, Kalythos$n_blind/Kalythos$total_sample,
       ylab = "proportion blind", xlab = "age")
lines(Kalythos$age, f0$fitted.values, col="red")
lines(nd$age,binomial()$linkinv(pred$fit), col = 4)
lines(nd$age,binomial()$linkinv(pred$fit+qnorm(0.05)*pred$se.fit), col = 4, lty = 2)
lines(nd$age,binomial()$linkinv(pred$fit+qnorm(0.95)*pred$se.fit), col = 4, lty = 2)
pred <- predict(f0, newdata = data.frame(age = c(20,50,70)),
                  se.fit=TRUE, type="link")
# confidence interval
cbind(binomial()$linkinv(pred$fit+qnorm(0.05)*pred$se.fit),
        binomial()$linkinv(pred$fit+qnorm(0.95)*pred$se.fit))
```

```
        [,1]      [,2]
1 0.08216859 0.1951734
2 0.56247042 0.6872294
3 0.83498665 0.9347823
```

```
# confidence interval width
binomial()$linkinv(pred$fit+qnorm(0.95)*pred$se.fit)-
      binomial()$linkinv(pred$fit+qnorm(0.05)*pred$se.fit)
```

```
        1         2         3
0.1130048 0.1247589 0.0997957
```



4. Si desidera testare:

$$H_0 : \beta_1 = 0.1 \; VS \; H_1 : \beta_1 \neq 0.1$$

Possiamo derivare un intervallo di confidenza o creare una statistica test ad-hoc:

```
confint.default(f0, parm = "age", level = 0.9) # 0.1 not in interval, reject H0
```

```
          5 %       95 %
age 0.0633402 0.09893884
```

```
 (tstat <- (f0$coefficients[2]-0.1)/summary(f0)$coef[2,2])
```

```
      age
-1.742916
```

```
 # pvalue
 2*pnorm(abs(tstat), lower.tail = FALSE)
```

```
       age
0.08134821
```

```
 # reject at 10%
 (tcrit = qnorm(.95))
```

```
[1] 1.644854
```

```
 # tstat in rejection region
 # reject H0
```

5. La stima per i parametri dei modelli GLM è derivata tramite la massimizzazione della verosimiglianza, quindi le stime derivate godono delle proprietà (asintotiche) degli stimatori di massima verosimiglianza: gli stimatori sono non distorti, hanno varianza nota derivata dalla matrice di informazione di Fisher e si distribuiscono approssimativamente secondo una normale. Il test derivato al punto 4 si basa su queste proprietà, ma la dimensionalità campionaria alla base delle stime non è particolarmente grande e si dovrebbe quindi essere cauti nell'applicazione di metodi basati sulle proprietà asintotiche dgli stimatori di massima verosimiglianza.

6.
```
 Kalythos$age_m20 <- Kalythos$age-20
 Kalythos$age_m45 <- Kalythos$age-45
 f20 <- glm(cbind(n_blind, n_healthy) ~ age_m20, data = Kalythos, family = binomial)
 f45 <- glm(cbind(n_blind, n_healthy) ~ age_m45, data = Kalythos, family = binomial)
 summary(f20); summary(f45) # same goodness of fit
```

```
Call:
glm(formula = cbind(n_blind, n_healthy) ~ age_m20, family = binomial,
    data = Kalythos)

Deviance Residuals:
      1        2        3        4        5
-0.1797   0.1157  -0.1182   0.3791  -0.3372

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.91499    0.30292  -6.322 2.58e-10 ***
```

```
age_m20      0.08114    0.01082   7.498 6.47e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 82.14455  on 4  degrees of freedom
Residual deviance:  0.31707  on 3  degrees of freedom
AIC: 24.132

Number of Fisher Scoring iterations: 4

Call:
glm(formula = cbind(n_blind, n_healthy) ~ age_m45, family = binomial,
    data = Kalythos)

Deviance Residuals:
      1        2        3        4        5
-0.1797   0.1157  -0.1182   0.3791  -0.3372

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.11350    0.15093   0.752    0.452
age_m45      0.08114    0.01082   7.498 6.47e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 82.14455  on 4  degrees of freedom
Residual deviance:  0.31707  on 3  degrees of freedom
AIC: 24.132

Number of Fisher Scoring iterations: 4

 predict(f0, newdata = data.frame(age = 20), type = "response")

        1
0.1284213

 predict(f20, newdata = data.frame(age_m20 = 20-20), type = "response")

        1
0.1284213

 predict(f45, newdata = data.frame(age_m45 = 20-45), type = "response")
```

```
        1
0.1284213
```

```
# same predictions
exp(f20$coefficients[1])/(1+exp(f20$coefficients[1]))
```

```
(Intercept)
  0.1284213
```

```
predict(f0, newdata = data.frame(age = 45), type = "link")
```

```
        1
0.1134985
```

```
f45$coefficients[1]
```

```
(Intercept)
  0.1134985
```

Le due intercette equivalgono ai valori di log(odd-ratio) per persone con 20 o 45 anni: i modelli stimati sono identici dal punto di vista della bontà di adattamento ai dati.

Esercizio 4

[Tratto da Salvan et al. *Modelli Lineari Generalizzati*, Springer]

Si prenda in esame il dataset contenuto nel file `bchem_phd.csv`. I dati contenuti nel data frame Biochemists (Long, 1990; Jackman, 2017) sono stati raccolti considerando dottori di ricerca in Biochimica che hanno conseguito il titolo nel periodo 1950-1967 in università degli Stati Uniti. Scopo dell'analisi era valutare le differenze di genere nella produttività scientifica. La variabile risposta è il numero di articoli scientifici, `art`, su riviste censite da Chemical Abstracts pubblicati nei 3 anni a cavallo del conseguimento del titolo. Le variabili concomitanti disponibili sono genere, `fem` (`Men`, `Women`), lo stato civile, `mar` (`Married`, `Single`), il numero di figli con non più di 5 anni, `kid5`, un indice di prestigio scientifico del dipartimento, `phd` (con valori tra 0 e 5), il numero di articoli scientifici pubblicati dal supervisore, `ment`, negli stessi 3 anni a cui è riferita la variabile `art`.

1. Si esamini la variabile risposta `art`.

2. Sarebbe possibile usare un modello di regressione multiplo per modellare la variabile risposta `art`? Quali vantaggi o svantaggi comporterebbe usare un modello di regressione multiplo per modellare la variabile risposta `art`?

3. Che modifiche si potrebbero apportare ad un modello di regressione multiplo per superare alcuni dei possibili svantaggi identificati al punto 2

4. Si usi un modello lineare generalizzato (GLM) in cui si assume che la variabile risposta `art` segua una distribuzione di Poisson per indagare se il numero di articoli pubblicati è influenzato dall'indice di prestigio scientifico del dipartimento, `phd`

5. Si trovi un sottoinsieme di predittori ottimali da usare in un modello lineare generalizzato simile a quello stimato al punto 3

6. Si produca una stima puntuale del valore atteso del numero di articoli per quattro persone con un dottorato in Biochimica con le seguenti caratteristiche:

```
nd <- data.frame(
    fem = c("Men","Women","Men","Women"), mar = c("Married","Married","Single","Single"),
    kid5=c(1,1,1,1), phh = c(3,3,3,3), ment = c(8,8,8,8))
rownames(nd) <- c("PHD A", "PHD B","PHD C","PHD D")
```

   Si commentino le stime trovate, esplicitando come i valori stimati dei coefficienti di regressione influiscono sui valori stimati

7. Si produca una stima intervallare usando un livello di confidenza del 90% per il valore atteso di articoli pubblicati dalle persone specificate al punto precedente

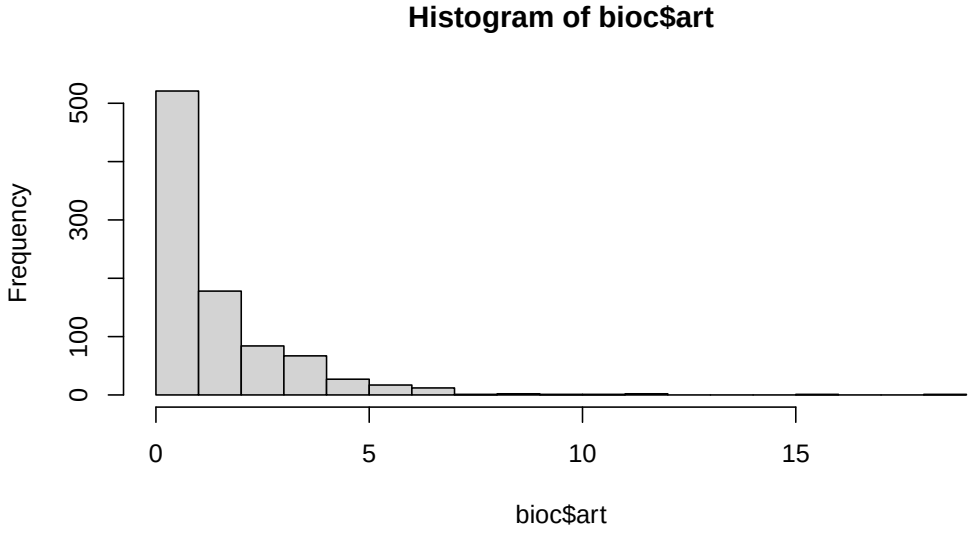
Soluzione Esercizio 4

```
bioc <- read.csv("data_exercises/bchem_phd.csv", header=TRUE)
```

1.
```
summary(bioc$art)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.000   1.000   1.693   2.000  19.000
```

```
hist(bioc$art,breaks = 20)
```



   La distribuzione è piuttosto asimmetrica con molti valori pari a 0 o comunque minori o uguali a 2. I valori della variabile sono discreti dato che corrispondono a dei dati di conteggio.

2. Il modello di regressione lineare è molto semplice da stimare e produce stime di coefficienti facili da interpretare, ma assume che la variabile risposta sia una normale, ma i dati in esame sono discreti e questo già è una violazione dell'assunzione di normalità. Sebbene i dati originali siano asimmetrici è possibile che i residui di un modello di regressione non risultino asimmetrici come la variabile originale. Tuttavia, se si usasse un modello di regressione multipla si potrebbero ottenere predizioni di valori negativi per certe combinazioni dei predittori: questo non ha senso a livello pratico dato che il numero di articoli può essere solo discreto.

3. Per evitare che si arrivi a stimare valori negativi per la variabile risposta si può pensare di modellare una trasformazione della risposta invece che le osservazioni originali: ad esempio si potrebbe costruire un modello per `log(art)` o `sqrt(art)`. Tuttavia il modo migliore per modellare questi dati è usare un GLM.

4. ```
   fit1 <- glm(art ~ phd, data  = bioc, family = poisson)
   summary(fit1)
   ```

   ```
   Call:
   glm(formula = art ~ phd, family = poisson, data = bioc)

   Deviance Residuals:
       Min       1Q   Median       3Q      Max
   -1.9596  -1.7599  -0.4767   0.3878   7.8155

   Coefficients:
               Estimate Std. Error z value Pr(>|z|)
   (Intercept)  0.25818    0.08666   2.979  0.00289 **
   phd          0.08532    0.02601   3.280  0.00104 **
   ---
   Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

   (Dispersion parameter for poisson family taken to be 1)

       Null deviance: 1817.4  on 914  degrees of freedom
   Residual deviance: 1806.6  on 913  degrees of freedom
   AIC: 3478.3

   Number of Fisher Scoring iterations: 5
   ```

   La variabile `phd` risulta significativa e positiva: all'aumentare del prestigio del dipartimento aumentano gli gli articoli.

5. Decidiamo di usare un algoritmo forward basato su AIC (altre scelte sono valide):

   ```
   selmod <- step(glm(art ~ 1, data  = bioc, family = poisson),
                  direction = "forward", trace = 0,
                  scope = list(upper = glm(art ~ ., data  = bioc, family = poisson)))
   summary(selmod)
   ```
   ```

```
Call:
glm(formula = art ~ ment + fem + kid5 + mar, family = poisson,
    data = bioc)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.6436  -1.5408  -0.3583   0.5623   5.3986

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.49735    0.05224   9.520  < 2e-16 ***
ment         0.02576    0.00195  13.212  < 2e-16 ***
femWomen    -0.22530    0.05461  -4.125 3.70e-05 ***
kid5        -0.18499    0.04014  -4.609 4.05e-06 ***
marSingle   -0.15218    0.06107  -2.492   0.0127 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1817.4  on 914  degrees of freedom
Residual deviance: 1634.6  on 910  degrees of freedom
AIC: 3312.3

Number of Fisher Scoring iterations: 5
```

6. ```
   nd <- data.frame(
       fem = c("Men","Women","Men","Women"), mar = c("Married","Married","Single","Single"),
       kid5=c(1,1,1,1), phh = c(3,3,3,3), ment = c(8,8,8,8))
   rownames(nd) <- c("PHD A", "PHD B","PHD C","PHD D")
   (pvals <- predict(selmod, newdata = nd, type = "response"))

     PHD A    PHD B    PHD C    PHD D
   1.679416 1.340635 1.442346 1.151388
   ```

   Il modello usa la funzione legame canonica, cioè il logaritmo. Il coefficiente stimato per l'effetto di essere donna è di -0.2253 e infatti i valori stimati per le donne sono il exp(-0.2253 = 0.7983) dei valori degli uomini: $0.7983 * 1.6794 = 1.341$ e $0.7983 * 1.4423 = 1.151$.

   Similmente, la stima per la variabile `mar` indica che le persone Single tendono ad avere in proporzione il 85.88 % di articoli in meno delle persone sposate (exp(-0.1522) = 0.8588): $0.8588 * 1.6794 = 1.442$ e $0.8588 * 1.3406 = 1.151$. .

7. ```
   pvals <- predict(selmod, newdata = nd, type = "link", se.fit = TRUE)
   cbind(selmod$family$linkinv(pvals$fit + qnorm(.05)*pvals$se.fit),
           selmod$family$linkinv(pvals$fit + qnorm(.95)*pvals$se.fit))
   ```

```
           [,1]     [,2]
PHD A 1.576407 1.789156
PHD B 1.231159 1.459846
PHD C 1.295342 1.606032
PHD D 1.029722 1.287429
```

Esercizio 5

[Esercizio di Esame aa 2018/2019 - prof. Gaetan]

Si consideri una ricerca sul comportamento dei clienti di un negozio online e sul rapporto tra vendite e apprezzamento del sito Web. A un certo numero di visitatori del sito Web è stato chiesto di esprimere il proprio apprezzamento per il sito Web su una scala Likert a 5 punti, che va da 1 (pessimo) a 5 (ottimo). Per questi visitatori è stato anche registrato se hanno effettivamente acquistato qualcosa sul sito web. I dati sono contenuti nel *file* `online.txt`.

1. Di che tipo sono le variabili coinvolte?
2. Si consideri un visitatore del sito Web e si supponga di non avere informazioni su quanto questo cliente apprezza il sito Web. Si stimi la probabilità che questo cliente abbia effettivamente acquistato qualcosa.
3. Si stimi la probabilità di acquistare qualcosa separatamente per ogni livello di apprezzamento e si mostri in un grafico le probabilità stimate in funzione dell'apprezzamento.
4. Si espliciti perchè non è sensato specificare un modello lineare per stabilire la relazione tra le due variabili.
5. Si stimi un opportuno modello di regressione logistica.
6. E' vero che la variabile *apprezzamento del sito* ha un effetto sulla probabilità d'acquisto? Se si, qual è quest'effetto? Quanto è forte l'evidenza a supporto della vostra affermazione?

Soluzione Esercizio 5

```
online <- read.table("data_exercises/online.txt", header = TRUE)
head(online)
```

```
  appreciation buy
1            4 yes
2            4 yes
3            3  no
4            4  no
5            2 yes
6            3 yes
```

1. `application` è una variabile di tipo ordinale (con 5 possibili valori), `buy` è una variabile dicotomica/categoriale con due possibili valori.
2. Senza avere ulteriori informazioni su un cliente la stima che si può usare è la media generale:

```
mean(online$buy == "yes")
```

```
[1] 0.5319149
```

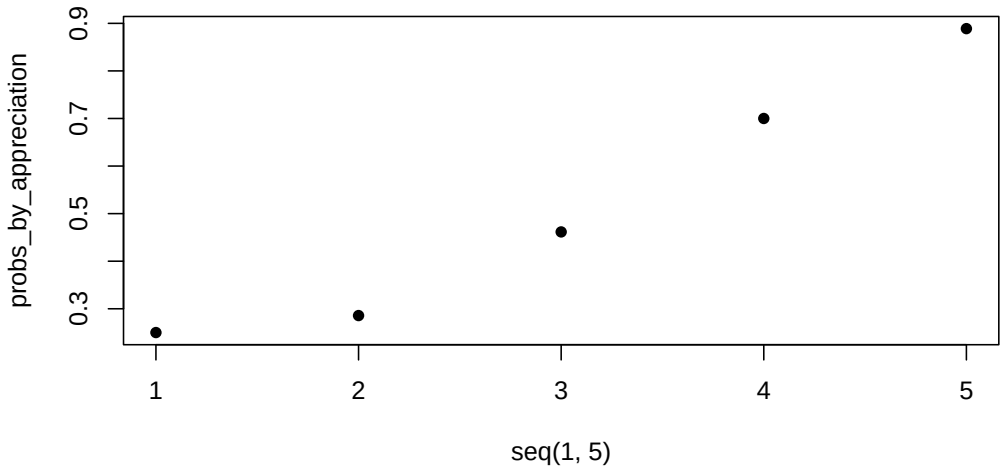
3.
```
(probs_by_appreciation <- tapply(online$buy == "yes",
                                 factor(online$appreciation), mean))
```

```
        1         2         3         4         5
0.2500000 0.2857143 0.4615385 0.7000000 0.8888889
```

```
# or by hand
mean(online$buy[online$appreciation == 1] == "yes") # etc
```

```
[1] 0.25
```

```
plot(seq(1,5), probs_by_appreciation, pch = 16)
```



Al crescere dell'apprezzamento cresce la probabilità che il cliente compri qualcosa

4. La relazione tra apprezzamento e probabilità di acquisto non è lineare: questo si potrebbe risolvere con delle trasformazioni, ma rimane il fatto che la variabile risposta è una proporzione, quindi qualcosa che deve essere per definizione in (0,1), mentre in un modello lineare la variabile risposta si assume normale e quindi definita sull'asse dei reali.

5. Ci sono due opzioni per stimare il modello: possiamo derivare il numero di osservazioni per ogni livello di `appreciation` o possiamo usare un modello bernoulli sui dati originali. Dobbiamo anche fare attenzione alla variabile esplicativa `appreciation` che può solo avere 5 valori, e va quindi trattata come un `factor`:

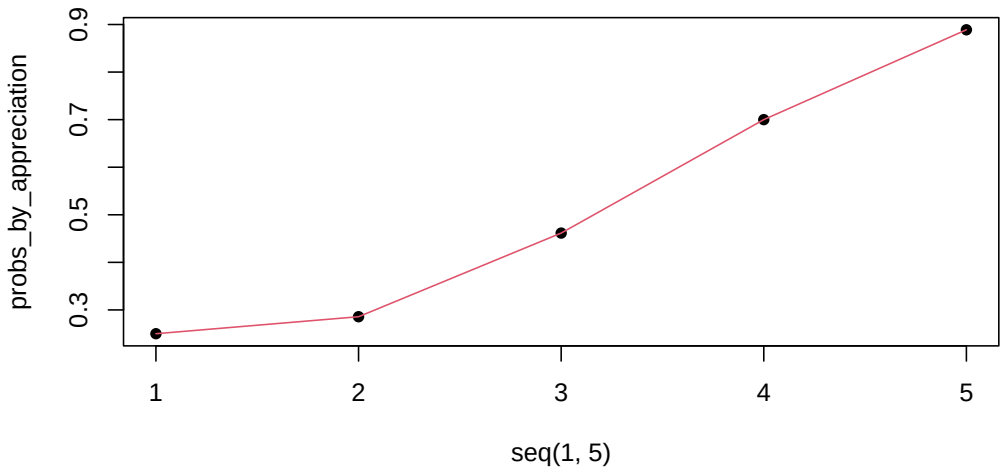
```
tot_obs_appreciation <- tapply(online$buy == "yes",
                               factor(online$appreciation), length)
fit_appreciation1 <- glm(probs_by_appreciation ~ factor(seq(1,5)),
                         weights = tot_obs_appreciation, family = binomial)
coef(fit_appreciation1)
```

```
      (Intercept) factor(seq(1, 5))2 factor(seq(1, 5))3 factor(seq(1, 5))4
       -1.0986123          0.1823216          0.9444616          1.9459101
factor(seq(1, 5))5
        3.1780538
```

```
fit_appreciation2 <- glm(factor(buy) ~ factor(appreciation), family = binomial(), data = onli
coef(fit_appreciation2)
```

```
         (Intercept) factor(appreciation)2 factor(appreciation)3
          -1.0986123            0.1823216            0.9444616
factor(appreciation)4 factor(appreciation)5
           1.9459101            3.1780538
```

```
plot(seq(1,5), probs_by_appreciation, pch = 16)
lines(seq(1,5), fitted(fit_appreciation1), col = 2)
```



6. Dato che abbiamo trattato `appreciation` come una variabile categoriale per verificare la significatività dell'inclusione della variabile possiamo usare la tabella della devianza e un likelihood ratio test:

```
null_model <- glm(probs_by_appreciation ~ 1,
                  weights = tot_obs_appreciation, family = binomial)
anova(null_model, fit_appreciation1, test = "LRT")
```

```
Analysis of Deviance Table

Model 1: probs_by_appreciation ~ 1
Model 2: probs_by_appreciation ~ factor(seq(1, 5))
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         4      11.15
2         0       0.00  4    11.15  0.02493 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

L'apprezzamento del sito ha un effetto sulla probabilità d'acquisto: dal grafico fatto al punto precedente si evince che l'effetto è monotono. Se si usasse `appreciation` come variabile continua si userebbero meno gradi di libertà (il modello `fit_appreciation1` usa tutti i gradi di libertà possibili) e si otterrebbe una stima meno discontinua dell'effetto di `appreciation`.

Esercizio 6

Il dataset `EdenRainfall` contiene informazioni sulle piogge registrate nel bacino del fioume Eden, nel nord dell'Inghilterra. Il dataset contiene le seguenti variabili:

- month: il mese a cui si riferisce l'osservazione
- year: l'anno a cui si riferisce l'osservazione
- ndays_prec: il numero di giorni con precipitazione $>$ 1mm nel mese
- high_prec: una variabile indicatore che ha valore 1 se nel mese è stata registrata una precipitazione estrema
- tot_prec: precipitazione totale accummulata nel mese
- mean_prec: precipitazione media accummulata nel mese
- tdays: il numero totale di giorni con records validi per il mese di riferimento
- nao: il valore del north atlantic oscillation (NAO) index per il mese.
- soi: il valore del southern oscillation index (SOI) per il mese.

Si indaghi se le variabili *nao* e *soi* influenzano la probabilità di osservare almeno un giorno con un elevata precipitazione (`high_precip`) e la proporzione di giorni piovosi in un mese per il mese di Gennaio. Si valutino in prima istanza modelli in cui i predittori vengono usati singolarmente: si creino grafici che mostrano l'impatto stimato dei singoli predittori sulla variabile di interesse. Si stimi un modello in cui i predittori sono entrambi inseriti nel modello. Si creino grafici che mostrano l'impatto dei predittori per diversi valori dell'altro predittore (ad esempio, 10o percentile, mediana e 90o percentile) sulla variabile di interesse.

Esercizio 7

Si prenda in esame la funzione `pois_gen_and_est` specificata nel seguente codice R:

```
pois_gen_and_est <- function(n, xrange = list(c(0,1)),
                             beta_true, out_est=FALSE){
   n_x <- length(xrange)
   X <- rep(1,n)
   for(j in 1:n_x) X <- cbind(X,
           runif(n, xrange[[j]][1], xrange[[j]][2]))
   y <- rpois(n, exp(X %*% beta_true))
   X <- X[,-1]
   if(!out_est) out <- data.frame(X,y)
   if(out_est) out <- as.numeric(coef(glm(y~X, family = poisson)))
   out
 }
```

1. Come vengono specificate le variabili risposta e i predittori? Che distribuzione ha la variabile risposta? Che distribuzione hanno i predittori? Si scriva in maniera estesa il modello sottostante la generazione dei dati nella funzione.

2. Si spieghi il contenuto dell'oggetto `pois_sim_n10` creato con il seguente codice:

```
NSIM <- 1000
set.seed(15496)
pois_sim_n10 <- t(replicate(NSIM,
            pois_gen_and_est(n=10, xrange = list(c(0,1), c(5,6)),
                             beta_true = c(1.2,1,0.6), out_est = TRUE)))
```

3. Si usi `pois_sim_n10` per quantificare lo standard error degli stimatori dei coefficienti di regressione in un GLM

4. Si usi la funzione `pois_gen_and_est` per studiare come lo standard error degli stimatori dei coefficienti di regressione in un GLM varia in funzione della dimensione del campione

5. Si usi la funzione `pois_gen_and_est` per studiare come lo standard error degli stimatori dei coefficienti di regressione in un GLM varia in funzione del vero valore dei parametri di regressione (nota bene: si consiglia di usare un modello con un solo predittore)

6. Si verifichi che quanto osservato al punto 5 sia in accordo con i risultati teorici presentati nelle slides

7. Si crei una funzione `binom_gen_and_est` per indagare il comportamento degli stimatori in un modello GLM per dati di tipo binomiale o bernoulliano.

8. Si usi la funzione `binom_gen_and_est` per indagare l'effetto di numerosità campionaria e vero valore dei coefficienti di regressione sull'incertezza degli stimatori